Data-centric Artificial Intelligence: A Survey

Year: 2023 | Citations: 306 | Authors: D. Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang

Abstract

Artificial Intelligence (AI) is making a profound impact in almost every domain. A vital enabler of its great success is the availability of abundant and high-quality data for building machine learning models. Recently, the role of data in AI has been significantly magnified, giving rise to the emerging concept of data-centric AI. The attention of researchers and practitioners has gradually shifted from advancing model design to enhancing the quality and quantity of the data. In this survey, we discuss the necessity of data-centric AI, followed by a holistic view of three general data-centric goals (training data development, inference data development, and data maintenance) and the representative methods. We also organize the existing literature from automation and collaboration perspectives, discuss the challenges, and tabulate the benchmarks for various tasks. We believe this is the first comprehensive survey that provides a global view of a spectrum of tasks across various stages of the data lifecycle. We hope it can help the readers efficiently grasp a broad picture of this field, and equip them with the techniques and further research ideas to systematically engineer data for building AI systems. A companion list of data-centric AI resources will be regularly updated on https://github.com/daochenzha/data-centric-AI.